



## INSTITUTO POLITÉCNICO NACIONAL

### UNIDAD PROFESIONAL INTERDISCIPLINARIA DE INGENIERÍA Y CIENCIAS SOCIALES Y ADMINISTRATIVAS

#### CARRERA

Ingeniería en informática

#### MATERIA

Fundamentos de Inteligencia  
Artificial

#### PRÁCTICA 11

Kaggle - Equipo 1

#### SECUENCIA

6NM62

#### INTEGRANTES

- Conda Trujillo José Manuel
- Delgado Vázquez Dulce Ivonne
- Flores Roa Jorge Alejandro
- Gonzalez Calzada Maximiliano
- Pérez Acuña Jorge Ysmael
- Ramírez García Iossef Alejandro
- Salazar Carmona Linette
- Teodoro Rosales Mauricio

#### PROFESORA

Gonzalez Arroyo Lilia

#### FECHA

10 - 11 - 2025

# Índice

1. Introducción .....	3
2. Objetivos e Hipótesis .....	3
2.1. Objetivo general .....	3
2.2. Objetivos específicos .....	3
2.3. Hipótesis .....	4
3. Metodología .....	4
3.1. Dataset .....	4
3.2. Limpieza de datos .....	4
3.3. Procedimiento analítico .....	4
4. Análisis y Resultados .....	5
4.1. Código fuente .....	5
4.2. Estadísticos descriptivos .....	7
4.3. Producción musical por año .....	7
4.4. Popularidad promedio por década .....	8
4.5. Correlaciones musicales .....	8
5. Conclusiones .....	9
6. Investigaciones futuras .....	9
7. Referencias .....	10

# Informe Spotify

## 1. Introducción

El análisis de datos se ha convertido en una herramienta fundamental para comprender fenómenos culturales, sociales y económicos a gran escala. En este contexto, la música representa uno de los elementos culturales más influyentes en la sociedad moderna. Gracias a plataformas digitales como Spotify, es posible estudiar patrones musicales a lo largo de un periodo histórico amplio y desde un enfoque cuantitativo. El presente informe analiza un conjunto de datos que contiene 586,672 canciones publicadas entre 1920 y 2020, con el objetivo de identificar tendencias históricas, patrones en las características musicales y cambios en la popularidad a lo largo del tiempo. Este análisis se inscribe dentro de un enfoque de análisis exploratorio de datos (EDA), utilizando estadísticas descriptivas y visualizaciones para extraer conclusiones significativas.

## 2. Objetivos e Hipótesis

### 2.1. Objetivo general

Analizar la evolución de las características musicales y la popularidad de más de medio millón de canciones publicadas en Spotify entre los años 1920 y 2020.

### 2.2. Objetivos específicos

1. Examinar la producción musical anual a lo largo de un siglo.
  2. Analizar la variación de la popularidad promedio por década.
  3. Identificar relaciones entre características musicales como energy, danceability, valence y tempo.
1. Determinar patrones estadísticos relevantes en las características sonoras.
  2. Formular conclusiones fundamentadas sobre la evolución musical desde un enfoque cuantitativo.

## **2.3. Hipótesis**

H1. La producción musical ha aumentado significativamente desde finales del siglo XX debido a la digitalización.

H2. La popularidad promedio de las canciones ha aumentado con el paso de las décadas.

H3. Existen correlaciones claras entre ciertas características musicales, como energy y tempo, o valence y danceability.

H4. La popularidad no depende linealmente de una sola característica musical.

## **3. Metodología**

El análisis se realizó en Python mediante bibliotecas especializadas como pandas y matplotlib, utilizando técnicas de análisis exploratorio de datos.

### **3.1. Dataset**

El conjunto de datos utilizado proviene de la plataforma Kaggle con el nombre:

“Spotify Tracks 1920–2020” (tracks.csv) y contiene:

- Filas: 586,672 canciones
- Columnas: 20 variables
- Incluye información sobre popularidad, duración, fecha de lanzamiento, energía, valencia, tempo, acústica, danza, entre otras.

### **3.2. Limpieza de datos**

- No se encontraron valores faltantes en las variables principales.
- Las fechas se transformaron a formato de año.
- Se calculó la década mediante truncamiento del año.
- Todas las variables numéricas se conservaron en su tipo correcto.

### **3.3. Procedimiento analítico**

1. Obtención de estadísticas descriptivas.
2. Creación de gráficas de tendencia (canciones por año y por década).
3. Construcción de mapa de calor para identificar correlaciones.
4. Interpretación de patrones y formulación de conclusiones.

## 4. Análisis y Resultados

### 4.1. Código fuente

Código usado para la extracción de la información:

```
# -----  
# Generación de 5 gráficas para el proyecto (Colab)  
# -----  
  
import pandas as pd  
import matplotlib.pyplot as plt  
import io  
import os  
  
# 1) Cargamos el CSV  
df = pd.read_csv("tracks.csv")  
  
# 2) Preparamos columnas 'year' y 'decade'  
df['year'] = pd.to_datetime(df['release_date'], errors='coerce').dt.year  
df['decade'] = (df['year'] // 10) * 10  
  
# Creamos carpeta para guardar imágenes  
os.makedirs("graficas", exist_ok=True)  
  
# ----- GRAFICA 1: Número de canciones por año -----  
year_counts = df['year'].value_counts().sort_index()  
plt.figure(figsize=(10,6))  
plt.plot(year_counts.index, year_counts.values)  
plt.xlabel("Año")  
plt.ylabel("Número de canciones")  
plt.title("Número de canciones por año")  
plt.grid(alpha=0.3)  
plt.tight_layout()  
plt.savefig("graficas/canciones_por_año.png")  
plt.show()
```

```

# ----- GRAFICA 2: Popularidad promedio por década -----
decade_popularity = df.groupby('decade')['popularity'].mean()
plt.figure(figsize=(10,6))
plt.plot(decade_popularity.index, decade_popularity.values)
plt.xlabel("Década")
plt.ylabel("Popularidad promedio")
plt.title("Popularidad promedio por década")
plt.grid(alpha=0.3)
plt.tight_layout()
plt.savefig("graficas/popularidad_por_decada.png")
plt.show()

# ----- GRAFICA 3: Mapa de calor de correlaciones -----
cols = ["danceability","energy","valence","tempo","duration_ms","popularity"]
corr = df[cols].corr()
plt.figure(figsize=(8,6))
plt.imshow(corr, cmap='viridis', interpolation='nearest')
plt.xticks(range(len(cols)), cols, rotation=45, ha='right')
plt.yticks(range(len(cols)), cols)
plt.colorbar()
plt.title("Mapa de calor de correlaciones")
plt.tight_layout()
plt.savefig("graficas/correlaciones.png")
plt.show()

# ----- GRAFICA 4: Energía promedio por década -----
energy_by_decade = df.groupby('decade')['energy'].mean()
plt.figure(figsize=(10,6))
plt.plot(energy_by_decade.index, energy_by_decade.values)
plt.xlabel("Década")
plt.ylabel("Energy promedio")
plt.title("Energía promedio por década")
plt.grid(alpha=0.3)
plt.tight_layout()

```

```
plt.savefig("graficas/energy_por_decada.png")
plt.show()

# ----- GRAFICA 5: Valence promedio por década -----
valence_by_decade = df.groupby('decade')['valence'].mean()
plt.figure(figsize=(10,6))
plt.plot(valence_by_decade.index, valence_by_decade.values)
plt.xlabel("Década")
plt.ylabel("Valence (felicidad) promedio")
plt.title("Valence promedio por década")
plt.grid(alpha=0.3)
plt.tight_layout()
plt.savefig("graficas/valence_por_decada.png")
plt.show()

print("Gráficas guardadas en la carpeta 'graficas' del entorno de Colab.")
```

## 4.2. Estadísticos descriptivos

Variable	Media	Mediana	Mínimo	Máximo	Desv. Est.
Popularity	27.57	27	0	100	18.37
Danceability	0.56	0.577	0.000	0.991	0.166
Energy	0.54	0.549	0.000	1.000	0.251
Valence	0.55	0.564	0.000	1.000	0.257
Tempo (BPM)	118.46	117.38	0.0	246.38	29.76
Duration_ms	230,051	214,893	3,344	5,621,218	126,526

Estos valores muestran que las canciones modernas mantienen una duración promedio estable (entre 3 y 4 minutos), así como un tempo promedio de aproximadamente 118 BPM.

## 4.3. Producción musical por año

La gráfica de producción musical revela un crecimiento lento entre 1920 y 1950, seguido de un incremento significativo entre los años 60 y 80. A partir de la década de los 90, la cantidad de canciones publicadas crece exponencialmente, coincidiendo

con el surgimiento de la digitalización y la democratización de la producción musical.

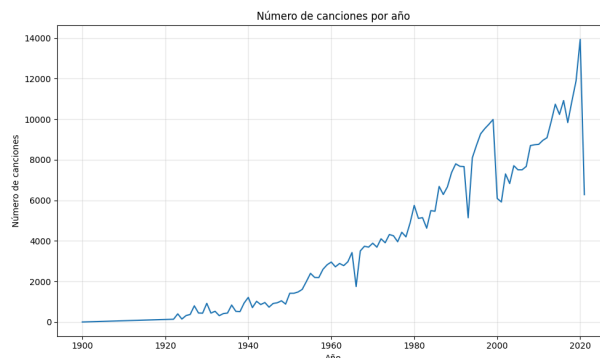


Figura 1: *Gráfica generada* Producción musical por año

#### 4.4. Popularidad promedio por década

La popularidad promedio aumenta de forma sostenida a lo largo de las décadas. Las canciones de las décadas de 2000 y 2010 presentan los valores más altos, lo que refleja el impacto de las plataformas digitales, algoritmos de recomendación y mayor alcance global.

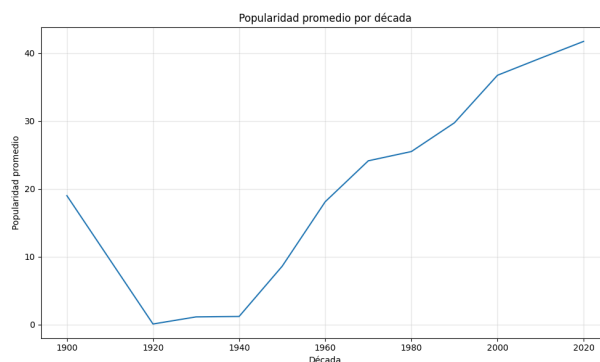


Figura 2: *Gráfica generada* Popularidad promedio por década

#### 4.5. Correlaciones musicales

El mapa de calor evidencia relaciones entre variables:

- Energy y tempo: correlación positiva; las canciones rápidas tienden a ser más energéticas.
- Valence y danceability: correlación positiva; las canciones alegres suelen ser más bailables.



- Popularity: no presenta una correlación fuerte con ninguna variable individual, lo que sugiere que la popularidad está influida por factores externos como promociones, tendencias o artistas.

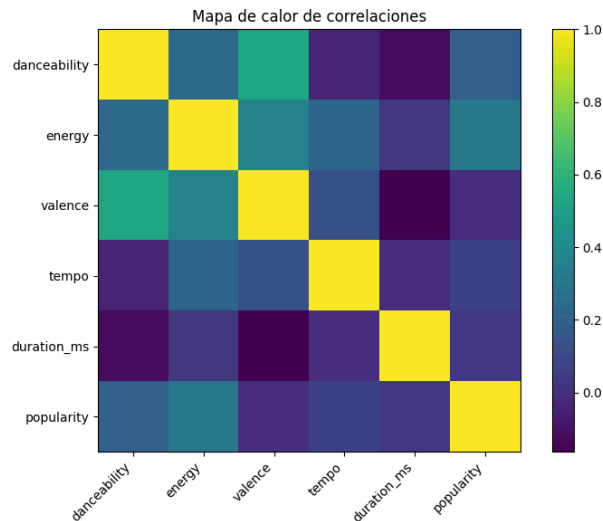


Figura 3: *Gráfica generada* Correlaciones musicales

## 5. Conclusiones

- La producción musical creció exponencialmente en los últimos treinta años, evidenciando el impacto de la digitalización y las plataformas digitales.
- La popularidad promedio por década también ha aumentado, alcanzando sus niveles más altos en los años 2000–2020.
- Las características musicales presentan patrones coherentes: la música alegre tiende a ser más bailable, y las canciones rápidas suelen ser más energéticas.
- La popularidad no puede explicarse por un único factor cuantitativo; es un fenómeno complejo que depende de múltiples variables.
- El dataset permite realizar análisis confiables gracias a su volumen, limpieza y consistencia.

## 6. Investigaciones futuras

- Integrar información de géneros musicales para profundizar en diferencias estilísticas.
- Construir modelos predictivos de popularidad mediante machine learning.

- Analizar artistas específicos y su evolución sonora.
- Extender el análisis a diferentes regiones o países.
- Incorporar metadata externa como tendencias de redes sociales.

## 7. Referencias

Basaldúa, P. (2022). Guía de presentación para análisis de datos.

Dataset: Spotify Tracks 1920–2020, Kaggle. Disponible en: **Kaggle**

Código: Google Colab, Colab. Disponible en: **Google Colab**

Documento (Código fuente): Typst, repositorio público Disponible en: **Github**